Research on E-mail Filtering Based On Improved Bayesian

Liu Pei-yu
（Shandong Normal University, Shandong Ji'Nan 250014, China）
E-mail: liupy@sdnu.edu.cn

Zhang Li-wei
（Shandong Normal University, Shandong Ji'Nan 250014, China）
E-mail: zgw-zlw@163.com

Zhu Zhen-fang
（Shandong Normal University, Shandong Ji'Nan 250014, China）
E-mail: zhuzhfyt@163.com

Abstract──**Naïve Bayesian has been widely used in spam filter because it simply and it also could classify texts more correctly and quickly. However, in the process of classifying and filtering, the traditional method doesn't consider the different features between the spam mail and the legitimate mail, and it also doesn't take into account the loss of misclassifying legitimate mail as spam, so there are many limitations of e-mail filtering. An improved algorithm based on Naïve Bayesian and Boosting method is proposed in this paper. The experiment result shows that the improved algorithm has better performance.**

Index terms─ **Spam; E-mail Filtering; Bayesian Algorithm; Boosting method**

I. INTRODUCTION

The popularity of the Internet added a new information channel of communication to people's [1]work, study and life. At the same time, the influence that it brought is more and more widespread. The emergence of large number of spam has become a social hazard. In order to enable Internet environment clean and beautiful, and make Internet users especially young students stay away from the harassment of non-friendly information, network information filtering technology has now become a hot research [1, 2].

With the rapid development of Internet, e-mail has become one of the most popular communicating modes for users for its convenience, speediness and cheapness. But spam (also referred to as "junk mail") is emerged with the convenience of e-mails, and bring harms to users. It is very late to begin researching on spam filtering in China; the spam in our country is quite serious. Nowadays, China has been the third most serious country in the world about the spam. So study of spam filtering is of great significance.

II. AN INTRODUCTION TO E-MAIL FILTERING

At present, E-mail filtering technology is the most used against the spam. E-mail filtering can be sub-divided into: Mail Transport Agent (MTA) filtering, Mail Delivery Agent (MDA) filters, Mail User Agent (MUA) form the filter architecture [3]. The main methods of anti-spam include based on E-mail address filtering technology, reversely verify domain names technology, based on rules (key words) filtering technology, and based on the statistical study of the E-mail filtering technology. Based on the statistical study of the E-mail filtering technology is the mainstream currently. In fact, based on the statistical study of the E-mail filtering technology is a binary classification. All kinds of text classification methods can be used for anti-spam, such as KNN Algorithm, SVM, Decision Tree and Bayesian Algorithm.

A. KNN Algorithm

KNN was proposed by Cover and Hart in 1968 at first, and theoretically it is a mutual approach. The thinking of the method is very simple and intuitive. We find k-nearest neighbors of a text (vector) in feature space. If most of them belong to a certain category, then this text (vector) belongs to the category, too. The advantages of this method are simple and accurate. When it comes to E-mail filtering, E-mails are divided into two types (legitimate E-mail and spam) in KNN Algorithm. We can find K nearest neighbors in the feature space. If majority of the K neighbors are belong to spam, the E-mail is classified as spam. The method is very simple, but its calculation is even larger, and can't adapt to the changing E-mails.

B. Support Vector Machine

Support Vector Machine (SVM) can be expressed as to find a hyper-plane, which can separate the data of the training set (in other words, it has minimum training

error) and has a minimum weight of vector. After using vector space model to express the characteristics of the document, the user's needs information can be seen as a document. That is to say, it can be expressed as a vector. The degree of similarity between documents and user's interests can be signified by the cosine similarity of the documents and user's interests.

SVM has the following features:

a. They have a minimum distance of the training sample from hyper-plane in all kinds of categories.

b. Though it has little vectors, it contains all the information which we used to classifying.

c. Most training samples are not support vector, so the removal or reduction of these samples has no impact on the classifier.

When the sample space can be non-linear sharing, the sample space can be mapped onto a high-dimensional space by using kernel function. Operational point plot can be computed in the new space. In order to facilitate calculation, we usually choose three kernel functions as mapping functions: polynomial kernel function, RBF and Sigmoid function. SVM is a structure based on risk minimization of the pattern recognition method.

When it is used to filtering E-mails, the method aims to find a hyper plane. This plane can clarify E-mails into legitimate E-mails and spam as rightly as possible, and the two types of the data are farthest form the plane. SVM can achieve good results of classification with a small training set. However, SVM can't adapt changing E-mails, too. Filtering effect is bad, when filtering new spam.

C. Decision Tree

In mathematics, the performances of the above methods are to build judgment surfaces. In fact, this surface is hard to build, because we will encounter various difficulties. Decision Tree is one of the most commonly method, because it has the following advantages:

a. As Decision Tree is divided into a few steps to carry out, the accuracy of the judgment is higher.

b. Judgment rules can be chosen some simple ones at each step.

c. we needn't to use all the features at each step. A few effective ones are enough. It will reduce the workload of each step.

d. classify faster.

The disadvantage of the method is that it will cost much time to build classifier. It has another question: the stability of the tree's structure. In addition, the use of the features and classification methods for different levels is a difficult job.

D. Bayesian Algorithm

"Bayesian" named after Thomas Bayes who was a famous mathematics, he developed a new possibility of inference theory which can predict the future by the past.

Bayesian classification model is a typical method of classification based on statistical models. Bayesian theorem is the most import formula in Bayesian theory, and it is the foundation to learning Bayesian. It combines the priori probability with the posterior probability wonderfully, and it predicts the posterior probability by using the prior probability and the sample data. Anti-spam model based on Bayesian Classifier deals with study of E-mails. It can be used to training E-mails sample set, identifying the division, refining the characters of the spam (training results). We can achieve Bayesian Classifier based on this model, use it to detect E-mail, and filter E-mail, if it is classified as spam. Bayesian filters are based on "self-learning" smart technology, and it can be adapt itself to the new tricks. Bayesian filters provider the legitimate E-mails with protection. Bayesian filtering technology has been applied to more and more anti-spam products, and it is one of the most effective anti-spam technologies. In spam filters, the naïve Bayesian has good effect, and it is very simple. An E-mail e_i can be expressed by $\vec{x}_{e_i} = (x_1, x_2, \ldots, x_n)$, and the probability of e_i belonging to C_k (generally, E-mails can be divided into two types, legitimate E-mail C_{legal}, and spam $C_{rubbish}$) is:

$$P(C_k \mid \vec{x}_{e_i}) = \frac{P(C_k)\prod_{l}^{n}P(x_l \mid C_k)}{\sum_{k \in (legal, rubbish)}P(C_k)P(\vec{x}_{e_i} \mid C_k)} \qquad (1).$$

In the above formula, we suppose that one feature is independent of the others, that is to say that the incidence of one word occurring in the E-mail is independent of the incidence of the other word occurring.

E. The Reason To Use Bayesian

Spam filtering and the generally classifications have many differences, as follows:

a. Both the mail severs and user client, the real-time requirements of the spam filtering is higher. We must choose the text classification which is simpler and faster as far as possible.

b. The relation of the general classification technology and text provider is either no matter or cooperative relations. But the relation of the spam manufacturers and filter developers is antagonistic relation. One side wants to filter spam as soon as possible. But the other wants to manufacture much spam and bypass the filters by all means.

c. When it comes to the results of the classification, people do not want to misjudge non-spam for spam. Otherwise, the non-spam will be filtered. It has nothing to do with that spam is misjudged as non-spam.

Compared with the above filtering algorithms, Bayesian filtering algorithm has the following advantages: (1) through the study of new spam and normal samples of spam, Bayesian will be able to combat the latest spam, and has better effect on various font. (2) Bayesian Algorithm is difficult to deceive. (3) In efficiency, Bayesian Algorithm is superior to the other algorithms. (4) Bayesian Algorithm can update with the constantly receiving E-mails. Therefore, Bayesian Algorithm is the most common method which is used to filtering spam, in the form of Naive Bayesian classifier.

III. IMPROVED NAIVE BAYESIAN ALGORITHM

We know that Naïve Bayesian algorithm is based on "Bayesian assumptions" which assume that each of the characteristics is independent by analyzing the theory. In fact, this assumption is difficult to exist. Experimental data also show that Bayesian algorithm makes important information lost, and leads to misjudging spam and legitimate mail. So, in this paper, we introduce Boosting method to the e-mail filtering field, and propose an improved filtering algorithm combined the Naïve Bayesian algorithm with Boosting method — the improved Naïve Bayesian algorithm based on Boosting method. This method's aim is to reduce the rate of the misjudgment, and to improve the accuracy of classification.

A. An Introduction To Boosting

Boosting method was proposed in the 1990s. Boosting is a general method which can boost the precision of the algorithm by many rounds learning and obtain much more accurate prediction rule by majority. This method can effectively transform the weak learning algorithm into strong learning algorithm. Boosting is a universal learning algorithm which can improve any given algorithm performance. Boosting has its roots in a theoretical framework for studying machine learning called the "PAC" learning model, due to Valiant [4]; see Kearns and Varian [5] for a good introduction to this model. Kearns and Valiant[6, 7] were the first to pose the question of whether a "weak" learning algorithm which performs just slightly better than random guessing in the PAC model can be "boosted" into an arbitrarily accurate "strong" learning algorithm. Sharpie [8] came up with the first provable polynomial-time boosting algorithm in 1989. A year later, Freund [9] developed a much more efficient boosting algorithm which, although optimal in a certain sense, nevertheless suffered from certain practical drawbacks. The first experiments with these early boosting algorithms were carried out by Drucker, Schapire and Simard[10] on an OCR task. Nowadays, Boosting has been simply and effectively used for modeling, image segmentation, intrusion detection, data mining and so on [11].

The basic idea of Boosting method is [12]: The algorithm takes as input a training set

$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ where each x_i belongs to some domain or instance space X, and each label y_i is in some label set Y. The algorithm calls a given weak or base learning algorithm repeatedly in a series of rounds $t = 1, \cdots, T$. The weight of this distribution on training example i on round t is denoted $D_t(i)$. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set.

The weak learner's job is to find a weak hypothesis h_t. Then the algorithm chooses a parameter $\beta_t (\beta_t \in C)$ which directly measures the importance of h_t. After T cycles, we can achieve the final assumption H by weighted voting. An individual weak learner has low rate of accuracy, but after we use boosting, the accuracy of the final outcome will be enhanced.

B. Improved Naïve Bayesian Algorithm Based On Boosting Method

We take the Naïve Bayesian algorithm as "weak" learning algorithm to be boosted. (In order to facilitate the narrative, we recorded the Naïve Bayesian as a filter function $h(x)$. There into, x stands for the e-mails to be classified.) According to the training example sequence (x_i, y_i) and the weight of distribution D_t, we get a filter function $h_t(x)$. At the same time, we update of each example's weight (give the larger weight to the incorrectly classified example). After T rounds of such training, we get a filter function sequence $h_1, h_2, \cdots, h_T$. Finally, we get a boosted filtering function $H(x)$ by the way of using a weighted majority vote.

The improved Naïve Bayesian algorithm based on Boosting method describes as follows:

Step 1. Input: sequence of N examples

$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ with $x_i \in X$ (X stands for training examples space) , $y_i \in C = \{c_1, c_2\}$ (c_1 stands for the class of legitimate mail, c_2 stands for the class of spam)

Initialize：

$$D_t(i) = \frac{1}{N}, (i = 1, \cdots, N) \qquad (2)$$

Step 2. Do for $t = 1, \cdots, T$ ：

（1）Train sequence of N examples inputted to minimize the filtering error of the example set construct $h_t(x)$ to minimize $D_t(i)[h_t(x_i) \neq y_i]$

（2）Calculate the error E_t（E_t is the sum of the weights of incorrectly classified examples）:

$$E_t = \sum_{t=1}^{T} D_t(i)[h_t(x_i) \neq y_i] \qquad (3),$$

Choose:

$$\beta_t = \frac{1}{2} \ln \frac{1-E_t}{E_t} \qquad (4)$$

（3）Update:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\beta_t y_i h_t(x_i))}{Z_t} \qquad (5)$$

, where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Step 3．Output：

$$H(x) = \sum_{t=1}^{T} \beta_t h_t(x) \qquad (6)$$

C. Evaluation Of The Algorithm

In the spam filter, we need to define some indicators to evaluate the spam filter effects in addition to the public corpus. These indicators are generally form text classification and information retrieval [13]. We assume that there are S(S=A+B+C+D) e-mails in the test set. To facilitate the narrative, we define some variables in Table I.

TABLE I: ADJACENT TABLE

	classified as legitimate mail by system	classified as spam by system
real legitimate mail	**A**	**B**
real spam	**C**	**D**

We define the evaluation indicators as follows:

a．Accuracy: Accuracy=$\dfrac{A+D}{S}$ *100%.

b．Recall: Recall=$\dfrac{D}{D+C}$ *100%. This indicator reflects the ability of filtering system to found out spam.

c．Precision: Precision=$\dfrac{D}{D+B}$ *100%. This indicator reflects the ability of filtering system to found spam correctly.

d．Miss Rate: Miss rate=$\dfrac{C}{D+C}$ *100%, the rate of spam which is not identified.

e．Error: Error=$\dfrac{B+C}{S}$ *100%，the rate of spam which is misclassified.

IV. EXPERIMENTAL RESULTS

We randomly selected 2000 e-mails from Chinese e-mail corpus of CCERT[14] and CASA[15] training examples to be classified, and randomly selected 400 e-mails as test set, including 100 legitimate mails and 300 spam. In order to prevent the chanciness of the experimental results, we have calculated the average of the results.

TABLE II: EXPERIMENTAL RESULT OF NAÏVE BAYESIAN ALGORITHM

	classified as legitimate mail by system	classified as spam by system	total
real legitimate e-mail	89.25	10.75	100
real spam	12.25	287.75	300
total	101.5	298.5	400

The experimental result of the improved Naïve Bayesian algorithm based on Boosting method is shown in table III.

TABLE III: EXPERIMENTAL RESULT OF IMPROVED FILTERING ALGORITHM

	classified as legitimate mail by system	classified as spam by system	total
real legitimate e-mail	96.5	3.5	100
real spam	6.33	293.67	300
total	99.83	300.17	400

Actually, there are 400 e-mails classified, including 100 legitimate mails and 300 spam. When the Naïve Bayesian algorithm is used, the system has classified 101.5 legitimate mails and 298.5 spam. The system has mistaken 10.75 legitimate mails as spam and has mistaken 12.25 spam as legitimate mail. And when we use the Improved Naïve Bayesian algorithm based on Boosting method, the system has classified 99.83 legitimate mails and 300.17 spam. The system has mistaken 3.5 legitimate mails as spam and has mistaken 6.33 spam as legitimate mail.

TABLE IV: COMPARISON OF EXPERIMENTAL RESULTS OF THE TWO FILTERING ALGORITHMS

	Naïve Bayesian	Improved filtering algorithm
Accuracy	94.25%	96.79%
Recall	95.92%	97.89%
Precision	96.4%	97.83%
Miss rate	4.08%	2.11%
Error	5.75%	3.21%

From Table IV, we know that the Improved Naïve Bayesian algorithm based on Boosting method has better performance than the traditional Naïve Bayesian algorithm in terms of Accuracy, Recall, Precision and so on. The improved filtering algorithm can not only improve the accuracy of spam filter, but also reduce the loss of the information and the error rate of misclassifying mail.

V. CONCLUSION AND FURTHER WORK

This article combined Bayesian algorithm with Boosting, presented a new spam filtering algorithm. Bayesian algorithm ignores the import information. The experiment shows that the algorithm can solve it. The algorithm is feasible. Our next step work begins with the following two aspects:

(1) In this paper, the feature vector we used is information gain. We will try other feature selected method, such as mutual information, TFIDF.

(2) Further study of the other advanced algorithms.

REFERENCES

[1] F. Crimmins, A. Smeaton, T. Dkaki, et al. Information discovery on the Internet. Intelligent Systems and Their Applications, IEEE. 1999: 55-62.

[2] Levitt, Mark Comiskey, Mike. Bright Light Focuses on Eliminating Spam. IDC Corporation. July 1998.

[3] Yong-jie Hu, Hong-xie Bo. Research on spam filtering technology. Journal of Hebei Normal University (Natural Science), 2006: 158-160.

[4] L. G. Valiant. A theory of the learnable. Communications of the ACM, 27(11):1134–1142, November 1984.

[5] Michael J. Kearns and Umesh V. Vazirani. *An* Introduction to Computational Learning Theory. MIT Press, 1994.

[6] Michael Kearns and Leslie G. Valiant. Learning Boolean formulae or finite automata is as hard as factoring. Technical Report TR-14-88, Harvard University Aiken Computation Laboratory, August 1988

[7] Michael Kearns and Leslie G. Valiant. Cryptographic limitations on learning Boolean formulae and finite automata. Journal of the Association for Computing Machinery, 41(1):67–95, January 1994

[8] Robert E. Schapire. The strength of weak learnability. Machine Learning, 5(2):197–227, 1990.

[9] Yoav Freund. Boosting a weak learning algorithm bymajority. Information and Computation, 121(2):256–285, 1995.

[10] Harris Drucker, Robert Schapire, and Patrice Simard. Boosting performance in neural networks. International Journal of Pattern Recognition and Artificial Intelligence, 7(4):705–719, 1993.

[11] FREUND Y, LYER R, SCHAPIRE R E, et al. An efficient boosting algorithm for combining preferences in machine learning. Proceedings of the Fifteenth International Conference. 1998:1-9.

[12] Schapire R E. A Brief Introduction to Boosting// Procedings of the 16th International Joint Conference on Artificial Intelligence, 1999.

[13] Zeng Chun, Xing Chunxiao, Zhou Lizhu. A Personalized Search Algorithm by Using Content-Based Filtering. Journal of Software[J]. 2003,14(5):999-1004.

[14] China Education and Research Network Computer Emergency Response Team. http://www.ccert.edu.cn/index.htm

[15] China Anti-Spam Alliance. http://www.antispam.org.cn/

Prof. Pei-yu Liu：Male, Broth in 1960, Doctoral tutor. Main research directions are computer network and information security, network system planning, network information resources development and software development technology.

Li-wei Zhang: Male, Broth in 1982, Postgraduate. Main research direction is information filtering.

Zhen-fang ZHU: Male, Broth in 1981, Postgraduate. Main research directions are information security and genetic algorithms.